Multi-armed Bandit을 이용한 요격 무장 할당 문제의 확률적인 접근*

홍택규○ 김건형 이병준 김기응
한국과학기술원 전산학부
{tghong, ghkim, bjlee}@ai.kaist.ac.kr, kekim@cs.kaist.ac.kr

A Probabilistic Approach to Interceptor Weapons Allocation Problem using Multi-armed Bandit

Teakgyu Hong○ Geonhyeong Kim Byung-Jun Lee Kee-Eung Kim
School of Computing, KAIST

요 약

본 논문에서는 적군이 아군의 기지를 향해 무장을 발사했을 때, 이를 요격하기 위해 발사할 요격 무장의 개수를 결정하는 문제를 다룬다. 기존의 요격 무장 할당 문제에 관한 연구들은 요격 무장의 요격 성공 확률을 알고 있다는 비현실적인 가정을 두었다. 하지만 실제 전쟁 중에는 상황에 따라 요격 성공 확률이 기존에 가정한 값과는 달라질 수 있으므로, 더욱 현실적인 연구가 되려면 이 확률이 알려져 있지 않다고 가정한 채 진행되어야 한다. 따라서 본 논문에서는 요격 성공 확률이 알려져 있지 않다는 가정을 바탕으로 요격 무장 할당 문제를 multi-armed bandit 문제로 모델링하여 이 문제를 해결하는 방법을 제시한다.

1. 서 론

본 논문은 적군이 아군의 기지를 향해 무장을 발사했을 때, 이를 요격하기 위해 발사할 요격 무장의 개수를 결정하는 문제를 다룬다. 이때 적군이 발사한 무장을 요격하기 위해 너무 많은 요격 무장을 발사하게 되면 아군의 자산을 낭비하게 되고, 너무 적은 요격 무장을 발사하게 되면 적군이 발사한 무장을 요격하지 못할 수도 있다. 따라서 요격 무장의 가격과 요격에 실패했을 때 받게 되는 피해를 고려해서 발사할 요격 무장의 개수를 효율적으로 결정하는 것은 중요한 문제이다.

요격 무장 할당 문제와 관련된 이전의 주요 연구로는 확률적인 최단 거리 문제(stochastic shortest path problem)를 이용한 연구[2]와 제약이 있는 최적화 문제(constrained optimization problem)를 이용한 연구[3]가 있다. 첫 번째 연구에서는 요격 무장 할당 문제를 상태(state)가 남아있는 아군의 요격 무장의 개수, 발사된 적군의 무장의 개수 등으로 구성된 확률적인 최단 거리 문제로 모델링한 후, 상태 집합의 크기가 큰 문제를 해결하기 위해 가치 함수(value function)를 인공 신경망을 이용하여 근사시켰다. 두 번째 연구에서는 요격 무장 할당 문제를 제약조건이 발사 가능한 요격 무장의 개수이고 최적화 문제가 요격 확률을 최대화하는 것인 제약이 있는 최적화 문제로 모델링한 후, 정수 계획법(integer programming)을 이용하여 발사할 요격 무장 개수를 계산하였다.

하지만 위의 두 연구는 모두 요격 무장의 요격 성공 확률을 알고 있다는 현실적이지 않은 가정을 하고 있다. 첫 번째 연구에서는 전투 시뮬레이터가 있다고 가정을 하고 있는데 이 전투 시뮬레이터에 요격 무장의 요격 성공 확률이 내재되어있다고 가정하고 있고, 두 번째 연구에서는 제약이 있는 최적화 문제를 모델링할 때 요격 무장의 요격 성공 확률을 알고 있다고 가정하고 있다. 그러나 실제 전쟁 중에는 상황에 따라 요격 성공 확률이 달라질 수 있으므로, 이 확률을 잘못 가정했을 경우 막대한 피해를 입을 수 있다.

본 논문에서는 실제적인 상황 고려한 요격 무장 할당 문제를 해결하기 위해 이 문제를 multi-armed bandit을 이용하여 모델링하였다. 그리고 multi-armed bandit 문제의 여러 알고리즘을 적용해서 요격 무장의 요격 확률을 모르는 상황에서도 발사할 요격 무장의 개수를 효율적으로 결정하는 방법을 제시하였다.

본 논문의 구성은 다음과 같다. 2장에서는 요격 무장 할당 문제를 모델링하기 위한 multi-armed bandit 문제에 대해서 알아보고, 3장에서는 모델링한 방법에 대해서 알아본다. 4장에서는 실험에 사용된 알고리즘들과 실험 결과에 대해서 알아보고, 마지막으로 5장에서 결론을 내리고 향후 연구방향을 제시한다.

2. Multi-armed Bandit 문제

Multi-armed bandit은 강화학습 문제의 한 갈래인 순차적 의사 결정(sequential decision making) 문제를 모델링하는데 사용되는 문제이다. 이 문제는 카지노에서 여러 대의 슬롯머신이 있을 때, 학습의 주체인 에이전트(agent)가 어느 슬롯머신을 선택해야 가장 높은 보상을 받을 수 있는지 직접 슬롯머신들을 선택해나가며 이를 알아내는 것이다[1]. 에이전트는 time step t마다 K개의 슬롯머신 중 하나를 선택하여 보상 r_t를 받는다. 본 논문에서는 각각의 슬롯머신을 선택했을 때 받게 되는 보상이 모수 p_i $(1 \le i \le K)$를 가지는 베르누이(Bernoulli) 분포로부터 나온다고 가정한다. 또한, 실험에서 베이지안(Bayesian) 방식의 알고리즘들도 사용하기

* 본 연구는 방위사업청과 국방과학연구소의 지원으로 한국과학기술원 초고속비행체특화센터에서 수행되었습니다.

위해 각각의 베르누이 분포의 모수 p_i는 모수가 a_i, b_i인 베타 분포를 따른다고 가정한다.

Multi-armed bandit 문제의 알고리즘들은 time step T때까지의 regret ρ값으로 성능평가를 하는데, 이는 T번째까지 최적 정책을 따라 슬롯머신을 선택하여 받은 보상의 합과 알고리즘을 따라 선택하여 받은 보상의 합의 차이의 기댓값이다. 이를 수식으로 나타내면 다음과 같고. 식에서 μ^*는 슬롯머신들의 보상의 기댓값 중 가장 큰 값이다.

$$\rho = T\mu^* - \sum_{t=1}^{T} r_t$$

3. 요격 무장 할당 문제의 모델링

본 논문에서는 발사 가능한 요격 무장의 개수를 최소 1개에서 최대 3개로 가정하고, i번째 슬롯머신을 선택하는 것을 요격 무장을 i개 발사하는 것으로 대응시켜서 요격 무장 할당 문제를 multi-armed bandit 문제로 모델링한다. 그리고 요격 무장을 1개 발사했을 때의 요격 성공 확률과 요격 무장을 여러 개 발사했을 때의 요격 성공 확률이 독립적인지 또는 종속적인지에 따라 두 가지의 경우로 나누어서 모델링하였다. 아래의 표 1에 이를 정리하여 나타내었다. 표 1에서 $\beta_{a,b}$는 모수가 a, b인 베타 분포를 의미하고, 사전 확률 분포는 요격 성공 확률에 대한 초기 확률 분포를 의미한다.

	독립적인 모델		
	1개	2개	3개
요격 성공 확률	p_1	p_2	p_3
사전 확률 분포	β_{a_1,b_1}	β_{a_2,b_2}	β_{a_3,b_3}
	종속적인 모델		
	1개	2개	3개
요격 성공 확률	p	$1-(1-p)^2$	$1-(1-p)^3$
사전 확률 분포	$\beta_{a,b}$		

표 1. 독립적인 모델과 종속적인 모델

요격 무장을 발사한 후 요격 성공 여부에 따라서 사전 확률 분포의 모수를 업데이트할 때, 독립적인 모델의 경우에는 단순히 요격 성공 여부에 따라 베타 분포의 모수를 증가시켜주면 된다. 하지만 종속적인 모델의 경우에는 독립적인 모델같이 단순한 방식으로 업데이트할 수 없다. 왜냐하면, 예를 들어 요격 무장을 2개 발사했을 때 요격에 성공했다면 2개 중 몇 개나 요격에 성공했는지 알 수 없기 때문이다.

본 논문에서는 다음과 같은 근사 방법으로 종속적인 모델에서 모수를 업데이트하였다. 예를 들어, 요격 무장을 2개 발사했을 때 요격에 성공했다고 가정하자. 그러면 그 때의 사후 확률 분포는 다음과 같다.

$$\begin{aligned}\beta_{a,b}\cdot(1-(1-p)^2) &= \frac{1}{B(a,b)}p^{a-1}(1-p)^{b-1}\cdot(p+p(1-p)) \\ &= \frac{1}{B(a,b)}p^{a}(1-p)^{b-1}+\frac{1}{B(a,b)}p^{a}(1-p)^{b} \\ &= \frac{B(a+1,b)}{B(a,b)}\beta_{a+1,b}+\frac{B(a+1,b+1)}{B(a,b)}\beta_{a+1,b+1}\end{aligned}$$

위의 식에서 알 수 있듯이, 사후 확률은 베타 혼합(beta mixture) 분포임을 알 수 있다. 본 논문에서는 모멘트 정합(moment matching) 기법을 이용하여 이 베타 혼합 분포를 새로운 하나의 베타 분포 $\beta_{a',b'}$로 근사하였다.

$$\mathrm{E}\left[\frac{B(a+1,b)}{B(a,b)}\beta_{a+1,b}+\frac{B(a+1,b+1)}{B(a,b)}\beta_{a+1,b+1}\right]=\mathrm{E}[\beta_{a',b'}]$$
$$\mathrm{var}\left[\frac{B(a+1,b)}{B(a,b)}\beta_{a+1,b}+\frac{B(a+1,b+1)}{B(a,b)}\beta_{a+1,b+1}\right]=\mathrm{var}[\beta_{a',b'}]$$

요격 무장을 3개 발사했을 때도 비슷한 방법으로 근사할 수 있다.

4. 실험

본 논문에서는 총 5가지 방법을 요격 무장 할당 문제를 모델링한 multi-armed bandit 문제에 적용했다.

1) 임의로 요격 성공 확률을 정하는 방법 (Fixed)
요격 성공 확률을 사전에 임의로 설정하고, time step마다 보상의 기댓값이 가장 높은 슬롯머신을 선택한다.

2) 최대우도 (Maximum likelihood) 추정법 (ML)
요격 성공 확률을 최대우도 추정법을 이용해서 추정하고, time step마다 추정한 요격 성공 확률을 바탕으로 보상의 기댓값이 가장 높은 슬롯머신을 선택한다.

3) Thompson sampling[4]
Time step마다 요격 성공 확률에 대한 사후 확률 분포에서부터 요격 성공 확률을 하나 샘플링한 후, 이 값을 바탕으로 보상의 기댓값이 가장 높은 슬롯머신을 선택한다.

4) Bayes-UCB[5]
Time step마다 요격 성공 확률을 사후 확률의 누적 분포 값이 $1-\frac{1}{t(\log n)^c}$이 되는 값으로 설정한다. 여기서 t는 현재 time step이고, n은 최대 time step이고, c는 상수로 본 논문에서는 이 값을 0으로 설정하였다. 설정한 요격 성공 확률을 바탕으로 보상의 기댓값이 가장 높은 슬롯머신을 선택한다.

5) KL-UCB[6]
Time step마다 요격 성공 확률을 아래의 값으로 설정하고 보상의 기댓값이 가장 높은 슬롯머신을 선택한다.

$$\max\left\{q\in\Theta : N[a]d\left(\frac{S[a]}{N[a]},q\right)\le \log t + c\log(\log(t))\right\}$$

여기서 Θ는 경계(bound)로 $\Theta=[0,1]$이고, $N[a]$와 $S[a]$는 각각 현재 time step t까지 총 슬롯머신 a를 선택한 횟수와 a를 선택했을 때 요격에 성공한 횟수이고, c는 상수로 본 논문에서는 이 값을 0으로 설정하였다. 그리고 $d(p,q)$는 베르누이 KL-발산(KL-divergence)으로 아래와 같이 정의된다.

$$d(p,q)=p\log\frac{p}{q}+(1-p)\log\frac{1-p}{1-q}$$

본 논문의 실험에서 요격 무장 당 가격은 -20으로, 요

격에 실패했을 때 받게 되는 피해는 -1000으로 설정하였다. 나머지 실험 환경 설정은 아래의 표 2에 나타내었다. 실험 환경은 각각의 모델에서 요격 성공 확률이 실제 값보다 과소평가 받고 있을 때와 과대평가 받고 있을 때의 두 가지 경우에 대해서 설정하였다.

독립적인 모델	과소평가			과대평가		
	p_1	p_2	p_3	p_1	p_2	p_3
실제 값	0.9	0.99	0.999	0.6	0.84	0.936
고정 값	0.6	0.84	0.936	0.9	0.99	0.999
ML의 초기 값	0.6	0.84	0.936	0.9	0.99	0.999
a_i, b_i의 초기 값	6.0, 4.0	8.4, 1.6	9.36, 0.64	9.0, 1.0	9.9, 0.1	9.99, 0.01

종속적인 모델	과소평가	과대평가
	p	p
실제 값	0.9	0.6
고정 값	0.6	0.9
ML의 초기 값	0.6	0.9
a, b의 초기 값	6.0, 4.0	9.0, 1.0

표 2. 실험 환경 설정

위의 표 2를 보면 독립적인 모델에서 무장을 1개 발사할 때의 확률 값과 여러 발 발사할 때의 확률 값이 서로 연관되어있는 것을 확인할 수 있다. 이는 독립적인 모델과 종속적인 모델 간의 비교를 위함이다. 아래의 그림 1과 그림 2는 time step마다의 regret을 나타낸 그래프이다.

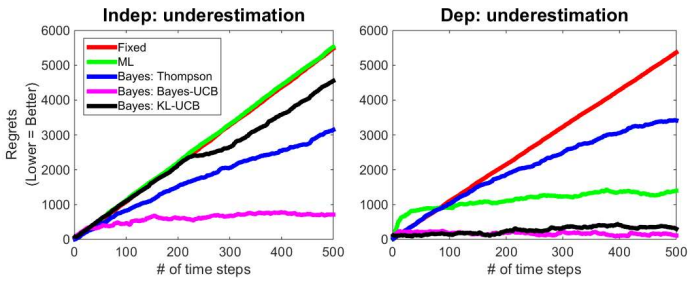

그림 1. 요격 무장의 요격 성공 확률을 과소평가할 때

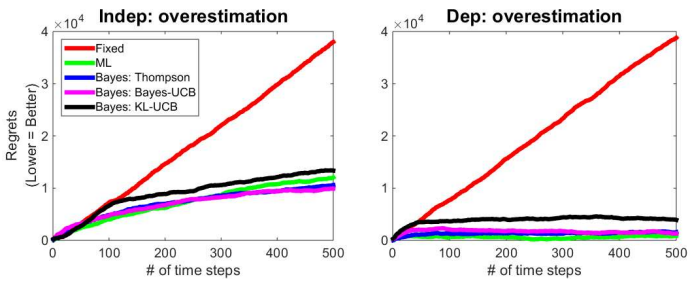

그림 2. 요격 무장의 요격 성공 확률을 과대평가할 때

실험 결과를 보면 전반적으로 독립적인 모델보다 종속적인 모델이 더 좋은 성능을 보임을 알 수 있다. 그 이유는 독립적인 모델은 슬롯머신마다 베타 분포의 모수를 독립적으로 가지고 있어서 정확한 요격 성공 확률을 추정하려면 각 슬롯머신을 여러 번 선택해봐야 한다. 하지만 종속적인 모델은 슬롯머신끼리 베타 분포의 모수를 공유하고 있으므로 독립적인 모델에 비해 적은 수의 시행을 하고도 요격 성공 확률을 보다 정확하게 추정할 수 있다. 그리고 실험에 적용한 여러 알고리즘들 중 본 논문의 실험 환경에서는 베이지안 방식의 알고리즘이 대부분 좋은 성능을 보였음을 알 수 있다.

5. 결론 및 향후 연구

본 논문은 적군이 아군을 향해 무장을 발사하였을 때, 이를 효율적으로 방어하기 위한 요격 무장의 개수를 결정하는 문제를 multi-armed bandit 문제로 모델링하여 해결하였다. 실험 결과 요격 무장을 한 개 발사했을 때와 여러 개 발사했을 때의 요격 성공 확률이 서로 연관되어있다고 가정한 모델이 더 좋은 성능을 보임을 알 수 있었고, 본 논문의 실험 환경에서는 베이지안 방법의 알고리즘이 대부분 좋은 성능을 보임을 알 수 있었다.

현재의 모델에서는 요격 무장의 요격 성공 확률을 시행착오를 통한 통계적인 방법으로 추론하고 있다. 하지만 실제 전쟁 상황에서는 기상 상황이나 무장들 간의 충돌 각도 등에 따라 요격 성공 확률이 달라질 수 있다. 또한, 아군의 남은 요격 무장의 개수 등의 제약 조건이 있을 수 있다. 따라서 지금의 모델을 contextual bandit이나 multi-objective bandit으로 확장하여 이러한 한계들을 극복하는 것이 향후 연구 방향이다.

참 고 문 헌

[1] Auer et al., "Finite-time Analysis of the Multiarmed Bandit Problem", in Machine Learning, vol. 47, 2002.

[2] Bertsekas et al., "Missile Defense and Interceptor Allocation by Neuro-Dynamic Programming", in IEEE Transactions on System, Man and Cybernetics, vol. 30, 2000.

[3] Karasakal, "Air defense missile-target allocation models for a naval task group", in Computers & Operations Research, vol. 35, 2008.

[4] Agrawal and Goyal, "Analysis of Thompson Sampling for the Multi-armed Bandit Problem", in Proceedings of COLT, 2012.

[5] Kaufmann, "On Bayes Upper Confidence Bounds for Bandit Problems", in Proceedings of AISTATS, 2012.

[6] Garivier and Cappé, "The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond", in Proceedings of COLT, 2011.